Visualization and analysis of RNA-Seq assembly graphs

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz599

Visualization and analysis of RNA-Seq assembly graphs

Fahmi W. Nazarie[1,†], Barbara Shih[1,†], Tim Angus[1], Mark W. Barnett[1], Sz-Hau Chen[1], Kim M. Summers[2,3], Karsten Klein[4], Geoffrey J. Faulkner[3], Harpreet K. Saini[5], Mick Watson[2], Stijn van Dongen[6], Anton J. Enright[7] and Tom C. Freeman[1,*]

[1]Systems Immunology Group, The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh EH25 9RG, UK, [2]Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh EH25 9RG, UK, [3]Mater Research Institute - The University of Queensland, Translational Research Institute, 37 Kent St, Woolloongabba QLD 4102, Australia, [4]Life Science Informatics Group, Department of Computer Science, Konstanz University, 78457 Konstanz, Germany, [5]Astex Pharmaceuticals, 436 Cambridge Science Park, Cambridge CB4 0QA, UK, [6]Cellular Genetics Informatics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA UK and [7]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK

Received December 19, 2018; Revised May 31, 2019; Editorial Decision June 27, 2019; Accepted July 10, 2019

ABSTRACT

RNA-Seq is a powerful transcriptome profiling technology enabling transcript discovery and quantification. Whilst most commonly used for gene-level quantification, the data can be used for the analysis of transcript isoforms. However, when the underlying transcript assemblies are complex, current visualization approaches can be limiting, with splicing events a challenge to interpret. Here, we report on the development of a graph-based visualization method as a complementary approach to understanding transcript diversity from short-read RNA-Seq data. Following the mapping of reads to a reference genome, a read-to-read comparison is performed on all reads mapping to a given gene, producing a weighted similarity matrix between reads. This is used to produce an *RNA assembly graph*, where nodes represent reads and edges similarity scores between them. The resulting graphs are visualized in 3D space to better appreciate their sometimes large and complex topology, with other information being overlaid on to nodes, e.g. transcript models. Here we demonstrate the utility of this approach, including the unusual structure of these graphs and how they can be used to identify issues in assembly, repetitive sequences within transcripts and splice variants. We believe this approach has the potential to significantly improve our understanding of transcript complexity.

INTRODUCTION

The advent of next generation sequencing platforms enables new approaches to solving a variety of problems in medicine, agriculture, evolution and the environment. RNA-sequencing (RNA-Seq) based transcriptome analyses are now used routinely as an alternative to microarrays for measuring transcript abundance, as well as offering the potential for gene and non-coding transcript discovery, splice variant and genome variance analyses (1,2). Data are typically summarized by counting the number of sequencing reads that map to genomic features of interest, e.g. genes. These measures are used as the basis for determining the level of expression in a given sample and differential expression between samples. Currently, a large number of pipelines for the analysis of RNA-Seq data have been developed to go from the output of a sequencing machine, to sequence assembly, and on to the quantification of gene expression (3–5). However, many aspects of the analysis of these data remain computationally expensive and limiting, and tools are still under active development (6–9).

One area in particular that remains challenging is the visualization and interpretation of transcript isoforms from short-read data. There are a currently a number of software tools available for the visualization of RNA-Seq assemblies, such as Manananggal (10), Integrative Genomics Viewer (IGV) (11,12), Tablet (13), BamView (14), EagleView (15), Artemis (16), Vials (17), SpliceViewer (18) and JunctionSeq (19) (reviewed in (20,21)). Perhaps the most widely used sequence analysis and visualization tool is the Integrative Genomics Viewer (IGV) (11,12). A distinguishing feature of IGV compared to other analysis/visualization

*To whom correspondence should be addressed. Tel: +44 131 651 9203; Email: tfreeman@roslin.ed.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

tools, e.g. Tablet (13), BamView (14), EagleView (15) and Artemis (16), is the inclusion of the Sashimi transcript visualization plot (22). Sashimi produces plots for RNA-Seq derived transcript isoforms providing a quantitative summary of genomic and splice junction mapping reads together with gene model annotations and read alignments. Alignments within exons are represented as read densities and paired reads connect exons, where the connection is weighted relative to the number of reads crossing a splice junction. Numerous other tools focus more on the quantitative and comparative visualization of differentially expressed spliced exons and isoforms in different samples (reviewed in (23)). In general, the majority of existing visualization approaches for splice variant analysis commonly involve the 'stacking' of reads onto a genomic reference. While this is sufficient for many needs, when the underlying transcript assemblies are complex and/or no reference genome is available, this approach can be limiting.

Novák *et al.* first introduced the idea of using graph-based methods to visualize DNA assemblies, in this case repetitive sequences in plant genomes (pea and soybean) (24). In this work, similarity scores between reads were precomputed by a series of computationally intensive pair-wise sequence alignments, and represented the first step in building a network from such data (24). After generating a matrix of similarity scores from an all-versus-all read comparison, read similarities exceeding a specified threshold were used to define network edges. In the visualization of the assemblies, nodes represented individual reads of DNA sequence and edges denoted sequence similarity, i.e. homology score between reads above a defined threshold. It was argued that the complex topology and diversity of the graphs produced could be used to better analyse the variability and evolutionary divergence of repeat families, as well as to discover and characterize novel elements. Graph visualizations were however generated as PDF files, limiting the opportunity for data exploration. By comparing the assembly-first approach to mapping-first approach in investigating alternative splicing, Benoit-Pilven *et al.* noted that the former enables more novel variants to be detected (25). Nielson *et al.* (26) developed an interactive network display called ABySS-Explorer that allows a user to display a sequence assembly and associated meta-data. ABySS was developed to assemble sequencing data derived from individual human genomes. Here, the de Bruijn graph data structure of *k*-mer neighbourhoods was used to reduce memory usage and computation time. Contig assemblies are represented as an oscillating line, the number of oscillations corresponding to a fixed number of nucleotides. However, the graphs were not designed to support the visualization of transcript splice forms. Bandage (27) is another a graph-based tool designed to visualize DNA assemblies following *de novo* assembly using de Bruijn graph methods (28). Here, contigs are shown depicted as long, flexible nodes whose length is proportion to the size of the contig, the connections (edges) between them being based on evidence of a likely association. This tool has been mostly used as a means of checking the assembly of bacterial genomes and is not well suited for the visualization of splice variants.

Here, we further explore the use of network-based visualizations of RNA-Seq data. Our aim has been to develop an approach that supports the visualization of DNA and RNA assemblies and provide a means to better understand transcript structure and splice-variation. In this paper, we develop a novel method for the visualization of RNA-Seq data using the graph analysis tool, Graphia Professional, formerly known as BioLayout *Express*^{3D} (29,30). In so doing, we provide a platform that supports the improved interpretation of complex transcript isoforms. We believe this approach will be useful in the exploration and discovery of new biological insights from sequence data.

MATERIALS AND METHODS

RNA-Seq data

Four samples of cell-cycle syncronized of RNA-Seq data were generated from serum-starved human fibroblasts (NHDF) (31). Briefly, cells were starved for 48 h and then harvested at 0 h and following serum refeeding at 12, 18 and 24 h, as the cells underwent synchronized cell division. RNA-Seq analysis was performed on an Illumina HiSeq2500 (Illumina, San Diego, California, USA) with 100 bp paired-end sequencing according to the manufacturer's recommendations and performed by Edinburgh Genomics (Edinburgh, UK) using the TruSeq™ RNA Sample Prep Kit (Illumina). Poly-(A) RNA was isolated and fragmented to produce of an average 180 bp fragments. Fragmented RNA was reverse transcribed and a single stranded DNA template was used to generate double strand cDNA which was blunt ended using T4 DNA polymerase prior to the addition of an adenosine base to assist ligation of the sequencing adapters. Flow cell preparation was carried out according to Illumina protocols; the libraries were denatured and diluted to a concentration of 15 pM for loading into the flow cells. RNA-Seq data were processed using the *Kraken* pipeline, a set of tools for quality control and analysis of high-throughput sequence data (32). Expression levels were reported as fragments per kilobase of transcript per million (FPKM).

To analyse a broader range of samples, RNA-Seq data from a human tissue atlas (33) representing 27 different tissues were downloaded from ArrayExpress database (E-MTAB-1733). Primary visualization of the data was performed using IGV to visualize the reads mapped on to the reference genome in certain loci or genes across samples. Long-read sequencing data of human heart, liver and lung samples released by Pacific Biosciences (PacBio) (34) were also utilized for making comparison to transcript assembly generated from the short read data.

Preparation of files for transcript visualization

The pipeline described below is based around a set of linked bash and Python scripts that perform the following tasks. Initial QC and read mapping to the reference genome (GRCh38) were performed using *BowTie* v1.1.0 (35). Sequence mapping data (BAM) were converted to a text file suitable for graph visualization in the free and open-source tool Graphia Professional https://kajeka.com/graphia-professional/ (Kajeka Ltd, Edinburgh, UK). Firstly, BAM files were sorted according to mapped chromosomal location using *sort* from *SAMtools* (36). The R

package *GenomicRanges* (37) was used to create annotation information, from a GTF file containing node annotation. This GTF file holds annotation information about gene structure (Ensembl version GRCh38). The output from this step was a tab-delimited file containing read mappings on Ensembl transcript and exon features. Exon junction spanning reads are assigned to the exon in which the majority of their sequence resides. This information can be overlaid on to graphs using the *class sets* function of Graphia, such that upon selection of an Ensembl transcript ID, nodes representing reads that map to this transcript model will be coloured according to the exon number.

The next step was to define the similarity between reads mapping to a gene of interest from the BAM and GTF files. A FASTA file containing all sequences mapping to a particular gene was extracted and the supporting information used for the visualization of transcript isoforms in the context of the resultant graph. For read-to-read comparison MegaBLAST (38) was used to generate a similarity matrix with edge weights derived from the alignment bit score. Parameterization of this step, i.e. defining the threshold for % sequence similarity (p) and length (l) over which two sequences must be similar in order for an edge to be drawn between them is of particular importance. Ideally, a graph should contain the maximum number of reads (nodes), connected by a minimum number of edges and where possible give rise to a single graph component, i.e. a single group of connected nodes that together represent the mRNA species of interest. For high coverage transcripts, more stringent parameters may be desirable.

Exploration of graph structure using simulated gene with multiple splicing events

Artificial transcript models representing two splice variants of the same 2706 bp gene were assembled from 10 exons of the gene *TTN*, selected exons being between 261 and 282 bp in length. When combined together, the two simulated transcripts incorporated an alternative start site (E1a, E1b), mutually exclusive exons (E3a, E3b), a skipped exon (E5) and an alternative 5′ donor site (20 bp shorter E7). Using ART (version MountRainier) (39) two levels of sequencing depth/transcript abundance were simulated, so as to provide either 250 or 1000, 125 bp reads per transcript model. For each level of transcript abundance, the simulated reads for the two transcripts were merged into a single FASTQ file and aligned to the reference genome (GRCh37) with HISAT (hierarchical indexing for spliced alignment of transcripts) (40). RNA assembly graphs were generated from the resulting BAM files using a percentage similar threshold ($p = 98$) and three settings for the threshold for length coverage ($l = 20, 40, 80$). The resultant graphs were visualized in Graphia Professional (Kajeka Ltd) (Figure 2).

Graph layout

The size and unusual topology of DNA/RNA sequence graphs necessitates the use of a highly optimized graph layout approach. Following experimentation (see Supplementary Figure S1), the Fast Multipole Multilevel Method (FMMM) (41) was shown to be well suited to the layout of these types of graphs. The FMMM algorithm was reimplemented in Java from the Open Graph Drawing Framework (OGDF) (42) and incorporated into the Graphia Professional code base (29), adding uniquely the ability to perform FMMM graph layout in 3D space. In general, the higher the FMMM quality setting, the more linear a graph becomes, but at the cost of computational runtime.

Collapsing of redundant reads

In the case of highly expressed genes, there can be a significant degree of redundancy in read coverage, i.e. reads of exactly the same sequence may be present in the data many times. Redundant reads add nothing to the interpretation of transcript structure and make the read-to-read comparison step unnecessarily time-consuming and the resultant graph sometimes difficult or impossible to visualize due to its size. Using *Tally* from the *Kraken* package (32), multiple identical reads were mapped to a single identifier that incorporates the number of occurrences of that specific sequence. When the read unification mode is employed, a single node is used to represent multiple identical reads, where the diameter of a node is proportional to the original number of reads it represents.

Analysis of the graph structure

Initially, we chose to examine a set of 550 genes whose expression was up-regulated as fibroblasts entered into S-M phases of the cell cycle (18–24 h after being refed serum). A graph derived from the 24 h data was plotted for each gene using MegaBLAST parameters $p = 98$, $l = 31$. Where the topology of a given gene graph was relatively simple, an explanation of its structure required only the overlay of individual transcript exon information in order to identify splice variant(s) represented. In other cases more detailed analyses were required. Other graphs were generated from the human tissue atlas data available at ArrayExpress (E-MTAB-1733) (33). In the tissue samples, reads may originate from multiple cell types expressing different isoforms of the same gene. The 100 bp paired-end reads for each tissue were individually mapped to the human genome (Ensembl GRCh38.82) with STAR v2.3.0 (43). The output from the mapping process (BAM files) was used to generate RNA assembly graphs using our pipeline. Publicly available data for *TPM1* were used to compare the network-based RNA-Seq approach with Pacific Biosciences (PacBio) long-read results obtained through their website (34). The *TPM1* gene models from both data were compared for heart, brain and liver.

Validation of splice variants using RT-PCR

To validate the existence of splice variants predicted by graph analyses, reverse transcription polymerase chain reaction (RT-PCR) of candidate splice variants was performed. Total RNA from human fibroblasts used for the RNA-Seq experiment was reverse transcribed in order to generate single stranded cDNA. Primers were designed using the Primer3 software (44) to amplify the region for validation of the splice variant. For *LRR1,* a pair of primers

was designed to amplify three splice variants as suggested from the graph visualization, while for *PCM1* two pairs of primers were designed across two different splice variant locations. For *LRR1*: Forward primer 5′-TGTTGAGCCTCT GTCAGCAG-3′ and reverse 5′-GTGTGGGCAACAGA ATGCAG-3′; for PCM1 (primer set 1) forward primer 5′-TCTGCTAATGTTGAGCGCCT-3′ and reverse 5′-TGCA GAGCTAGAAGTGCAGC-3′ and *PCM1*: (primer set 2) Forward 5′-ACGGAAGAAGACGCCAGTTT-3′ and reverse 5′-AGCTGCAGCTCATGGAAGAG-3′. PCR was carried out for 35 cycles (92°C, 30 s; 60°C, 90 s; 72°C, 60 s). The amplicons were run on a 2% agarose gel in the presence of SYBR-Safe DNA gel stain (Thermo Fisher, Waltham, MA, USA) and gels visualized by UV illumination.

Access to the pipeline and web-based NGS Graph generator

Documentation and full source code for the NGS Graph Generator package can be downloaded from: https://github.com/systems-immunology-roslin-institute/ngs-graph-generator. The user needs to supply BAM and GTF files to run the pipeline. In addition, we have developed a web interface designed for demonstration purposes rather than the analysis of a user's own data, that allows the pipeline to be run on a number of predefined datasets. This web-interface is called *NGS graph generator* and can be accessed at http://seq-graph.roslin.ed.ac.uk. Using this resource, a user can select a BAM file from RNA-Seq time-course samples of human fibroblasts or data from the human tissue atlas. Users can adjust parameters used by MegaBLAST to compute read similarity and there is an option to discard identical reads. Processing time required is dependent on the number of reads mapping to a gene of interest. The user must provide their email address and will be informed once the job finished. The resultant graph layout file will automatically open Graphia Professional (if installed). Protocols for graph generation and visualization are provided in Supplementary File S1 and a video of graph visualizations is provided in Supplementary File S2.

RESULTS

The principle of rendering sequence data as a network is illustrated in Figure 1A. We have developed a pipeline to create such graphs from RNA-Seq data, where the outputs can be visualized as a graph within Graphia Professional (Figure 1B). For these studies, we generated RNA-Seq data from four samples of human fibroblasts, taken at different time points during synchronized cell division. Paired-end cDNA libraries for each RNA sample were prepared and sequenced (see Materials and Methods). Other RNA-Seq data used for these studies originated from a publicly available human tissue atlas experiment (33) (E-MTAB-1733) which was also processed through the pipeline.

Simulated transcript analysis

Simulated RNA-seq data was generated for two synthetic transcript isoforms, such that when the isoforms were combined it generated a model of a single gene loci exhibiting the four main splicing event types: (a) an alternative start site; (b) mutually exclusive exons; (c) exon skipping event and (d) alternative splice site (Figure 2). To simulate the effect of using different similarity thresholds for sequence comparison, the length of the region showing sequence similarity was varied ($l = 20, 40, 80$). Another factor that can be adjusted in defining the comparison threshold is the percent identity (p) over the length (l), but for these studies p was maintained at 98%. The influence of transcript abundance was also tested and read depths of either 500 or 2000 reads per assembly were employed. Graphs were then constructed for each condition. Varying the stringency for read similarity thresholds ($l = 20$ or 40) for the low read depth simulated data, resulted RNA assembly graphs that were comprised of a single connected component and three of the four (a–c) splicing events were clearly visible. At the highest stringency setting ($l = 80$), the graph fragmented into four components and as such, would be very difficult to interpret. With the higher read depth data, the graphs generated at all read similarity thresholds existed as a single connected component. Splice events (a–c) were clearly visible at all threshold settings, but only at the highest stringency setting ($l = 80$) was the 20 bp 5′ alternative splice site in exon 7 visible. These studies demonstrate that if the thresholds used are too stringent, graphs may fragment, if too low, the underlying structure may be obscured. However, when read depth is high, even small changes in the structure of transcripts can be observed. After empirical exploration of the two variables for thresholding the read-to-read similarity score using a range of RNA assembly graphs, we chose a percentage similarity $p = 98$ and percentage length coverage $l = 31$, as reasonable generalized values for these variables.

Graph complexity reduction

In some instances where the level of expression is especially high, graph visualization is not possible due to the number of nodes and edges needed to represent the data. For instance, in the 24 h serum-refed fibroblast samples the highly expressed genes *TUBA1C* and *GAPDH* had 38,294 and 59,998 reads mapping to them, respectively. Node reduction is a process whereby identical reads are collapsed down to and represented by a single node, the size of the node being proportional to the number of reads it represents. In the case of *TUBA1C*, this reduced the number of nodes from 38,294 to 6511 (Figure 3A), whilst the number of edges was reduced from 90,340,179 to 1,779,069. In the case of *GAPDH*, the reduction in nodes was from 59,998 to 9264, whilst the number of edges was reduced from 208,221,932 to 3,562,688 (Figure 3B). The reduced graph for *GAPDH* is also shown generated at two different MegaBLAST threshold settings. On the left the graph was generated at the default BLAST setting of $p = 98$, $l = 31$, the second used a more stringent setting of $P = 98$, $l = 95$. Such is the depth of sequencing of this gene that even using a BLAST setting of 98% similarity over 95 bp of length the graph still forms a single component where the number of edges is reduced by approximately 90% but the number of nodes by only ∼1%. At this higher stringency BLAST setting, the graph uncoils

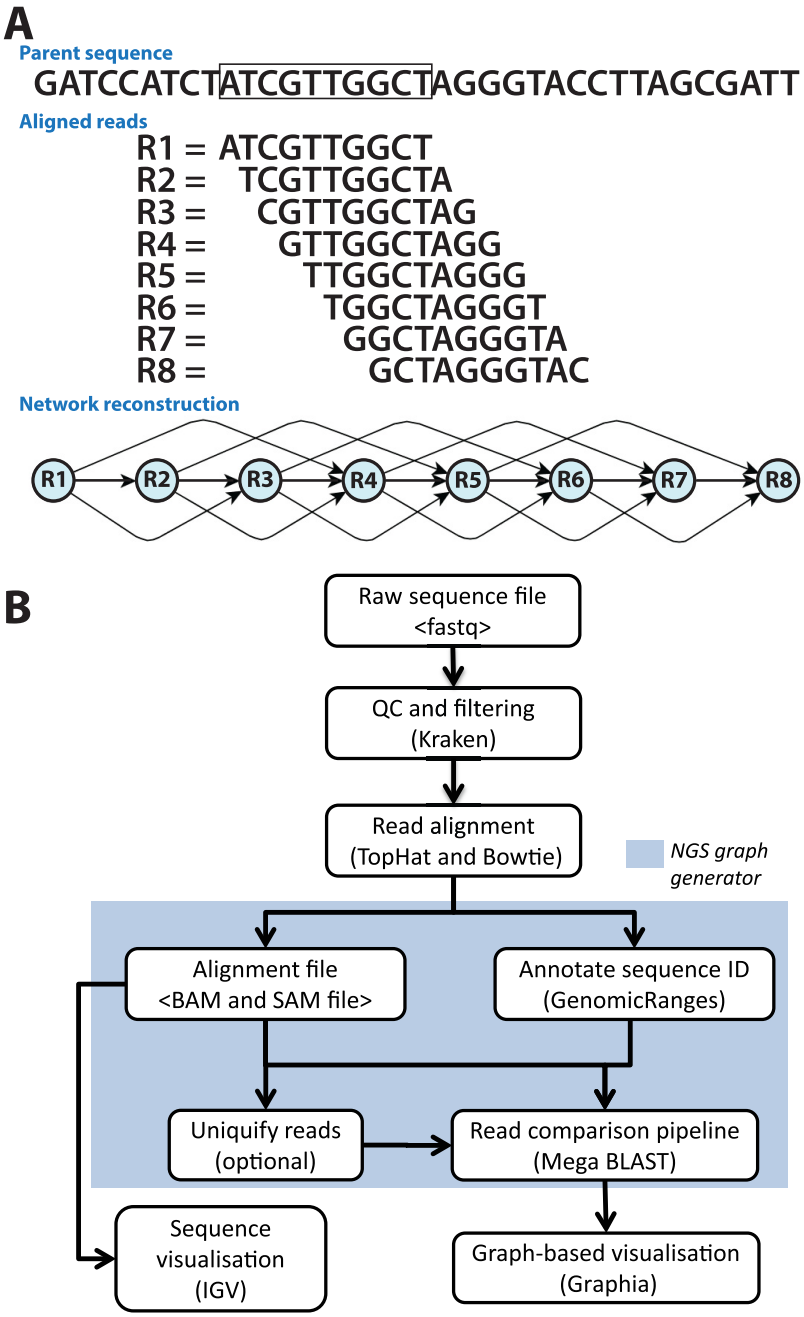

Figure 1. Development of the analysis pipeline for visualization of RNA-Seq data. (**A**) Network paradigm for the analysis of short-read RNA-Seq data. A region of DNA is shown with 10 bp 'reads' aligned to it below. In this view, eight different reads position with a 1 bp consecutive offset to the parent sequence. If reads were compared and a threshold of ≥70% similarity were used to construct a graph, the network would have structure shown. (**B**) Pipeline for network-based visualization of RNA-Seq data. Analysis pipeline for building graphs from RNA-Seq data passes from raw sequencing FASTQ file through a series of analysis/annotation steps up to the production of a file for graph visualization within Graphia Professional.

exposing small nodes representing unique reads due to sequencing errors (as shown in inset of Figure 3B).

Network visualization of transcripts

Having explored the settings for network construction, we individually examined 550 RNA assembly graphs from the set of genes regulated as NHDF cells undergo mitosis (31). In the majority of cases, the graphs for these genes were linear, i.e. possessed no higher order structure; two example transcripts, *KRT19* and *CCNB1*, from two time points (0 h and 24 h post-serum) are shown in Figure 4. Arguably little has been learnt by visualizing these data in this manner. However, even with these networks transcript variance could be observed. The change in expression level can be inferred from a reduction/increase in the number of nodes. In the network representing *CCNB1* there was also clear evidence that two transcript isoforms were expressed by the fibroblasts, one of which was truncated at the 3′ end.

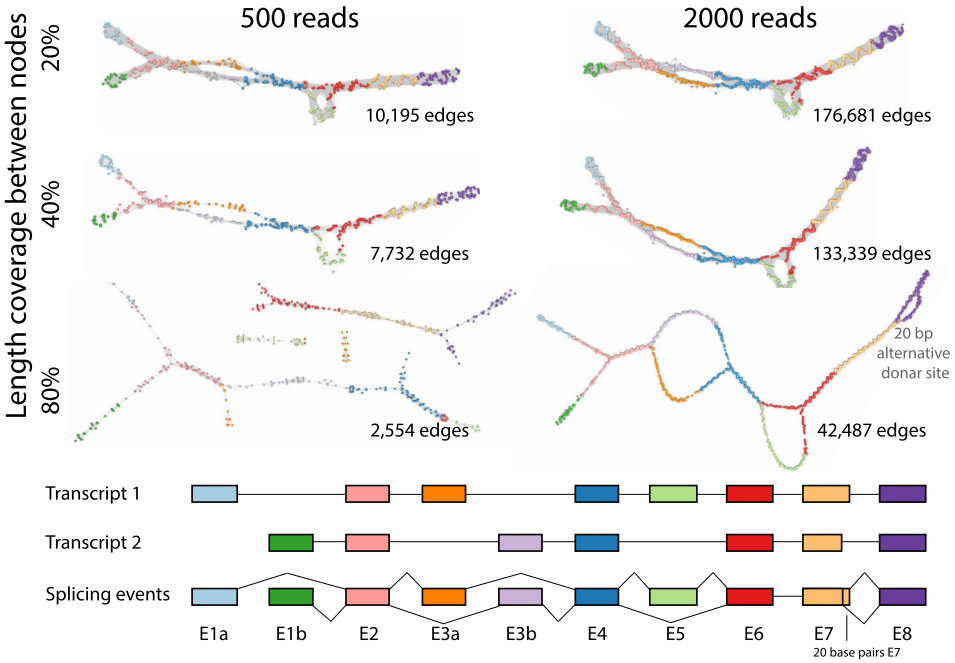

Figure 2. Graph visualization of simulated data. Artificial transcript models representing two splice variants of a 10 exon gene were generated (bottom). When combined together, the two simulated transcripts incorporated an alternative start site (E1a, E1b), mutually exclusive exons (E3a, E3b), a skipped exon (E5) and an alternative 5′ donor site (20 bp shorter E7). Two levels of sequencing depth/transcript abundance were simulated, so as to provide a total of 500 or 2000 125 bp reads for the combined transcript models. RNA assembly graphs were then generated using a percentage similarity threshold of $p = 98$ and three settings for the threshold for the length of similarity ($l = 20, 40, 80$). The resultant graphs were visualized Graphia Professional using a 2D layout. Graphs are laid out in 2D and node colours represent the exon to which they map.

This was manifest by the fact that there were fewer reads present at the 3′ end of the network, the falloff in read density occurring at the point where a known variant occurs (ENST00000505500) (Figure 4iii). This decrease in reads at the 3′ end of *CCNB1* is also visible in the standard IGV Sashimi plot (Figure 4iii).

Splice variant network structure

Of the 550 differentially expressed transcripts examined in the NHDF data, approximately 5% of RNA assembly graphs exhibited complex topologies. We investigated the underlying reasons for these non-linear structures. LRR1 (leucine-rich repeat protein 1) is known to regulate the cell cycle in *Caenorhabditis elegans* and actin-based motility in human cells (45). The *LRR1* graph comprised of a single loop-like structure and corresponded to the two known transcript isoforms for this gene. One (ENST00000318317) has only three exons, where exons 3 and 4 are skipped, while a second protein coding transcript (ENST00000298288) contains four exons, numbered 1, 2, 4 and 5 (Figure 5A). There was also evidence for the presence of a nonsense mediated decay product (ENST00000554869) as a small number of reads partially mapped to a small exon (numbered exon 3) specific to this transcript. *PCM1* (pericentriolar material 1) is a 6,075 nt gene containing 36 exons that encodes a protein that recruits PLK1 to the pericentriolar matrix to promote primary cilia disassembly before mitotic entry (46). There are seven known protein coding variants of PCM1 and network analysis provided evidence for two splicing events when expressed in fibroblasts (Figure 5B). One loop was indicative of the splicing out of exon 7 and the other of exon 24, indicating the presence of transcripts ENST00000517730 and ENST00000522275 respectively, in addition to the main isoform of this gene (ENST00000325083). RT-PCR confirmed the splicing events for *LRR1* and *PCM1* genes predicted by the network-based analysis (Figure 5Aiv and Bii). The visualization of splice variants was also supported in the Sashimi plots for these genes, but even in these relatively simple examples of splice variation the plots can be challenging to interpret, especially in the case of *LRR1* (Figure 5Aiii).

Observations of issues with assembly and internal repeats

CENPO (Centromere O protein) is a component of the Interphase Centromere Complex (ICEN) and localizes at the centromere throughout the cell cycle (47), where it is required for bipolar spindle assembly, chromosome segregation and checkpoint signalling during mitosis. The RNA assembly graph of *CENPO* showed a complex topology within its final 3′ exon (Figure 6A). In principle, network elements representing single exons should form linear graphs with bifurcations from linearity only occurring at exon junctions. In order to explain the observed anomalies in the graph for this gene, we investigated the genomic origin of reads mapping to loop junctions using BLAST. It transpired that many mapped to exon-exon junctions of adenylate cyclase 3 (*ADCY3*), a gene located on the opposite strand of chromosome 2. A number of its 5′ exons overlap with the final 3′ exon of *CENPO*. The RNA-Seq libraries

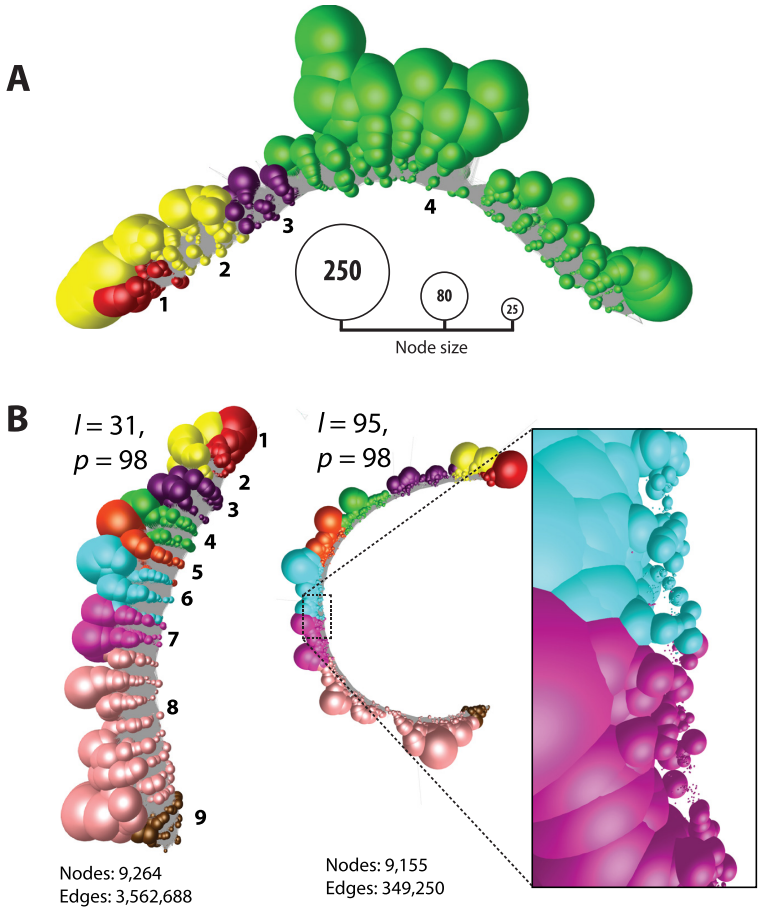

Figure 3. Read uniquification of highly expressed genes *TUBA1C* and *GAPDH* in human fibroblasts. (**A**) Graph representation of *TUBA1C* after read uniquification. The number of reads mapping to *TUBA1C* was 38,294 in the fibroblast sample. After read uniquification this was reduced to 6511 unique reads. When graphs are collapsed down to unique sequences node size is proportional to the number of individual reads represented, and nodes have been coloured according to the exon on to which they map. The *TUBA1C* transcript model matching the graph here is ENST00000301072, a 3,001 bp transcript encoding a 449 amino acid protein. (**B**) Graph for *GAPDH* after read uniquification as described above but shown at two different thresholds of BLAST scores. On the left using our default threshold of 98% similarity (*p*) over 31 bp (*l*), on the right 98% similarity over 95 bp. Increasing the edge threshold (*l*) dramatically reduces the number of edges, whist the number of nodes is barely affected. It also opens up the structure and when edges are removed the large number of small nodes representing unique reads due to sequencing errors can be observed (inset). The *GAPDH* transcript model matching the graph here is ENST00000396856, a 1266 bp transcript encoding a 260 amino acid protein.

were non-directional in nature and due to an ambiguity in read mapping, reads in the assembly of *CENPO* were actually derived from *ADCY3*. Reads from exon boundaries of *ADCY3* gave rise to the observed alternative splicing-like structures within the final portion of the *CENPO* graph. These anomalies are difficult to observe using conventional visualization tools such as IGV and even with the Sashimi plot it is not easy to distinguish which of the two overlapping genes reads are derived from.

MKI67 encodes antigen Ki-67, a well-established cell proliferation marker, that helps support the architecture of the mitotic chromosome (48). The graph of *MKI67* contains two features; a loop representing a known splice variant, and a knotted structure associated with exon 14. In the case of the former, exon 7 is spliced out in transcript ENST00000368653 as compared to ENST00000368654, both isoforms being expressed within fibroblasts. In the second case there are 13 repeats of a K167/chmadrin domain within exon 14, the internal homology leading to the formation of the observed structural complexity (Figure 6B).

Human tissue graph analysis of *TPM1* gene

In order to explore transcript variation within and between tissues we examined expression patterns across a human tissue atlas. We focus here on *TPM1*, a gene that encodes the muscle/cytoskeletal protein alpha tropomyosin. This gene was selected because it is widely expressed across tissues but at very variable levels, and was identified by the *rMATS* package (49) as having a high statistical probability of expressing multiple splice variants between the tissues examined, i.e. heart, liver, brain. *TPM1* has 15 exons and Ensembl reports there to be 19 protein coding transcript isoforms, with a further 14 non-coding isoforms, i.e. with a retained intron or the products of nonsense mediated decay. We analysed graphs of *TPM1* generated from three different human tissues (heart – 10,347 reads; liver – 431 reads;

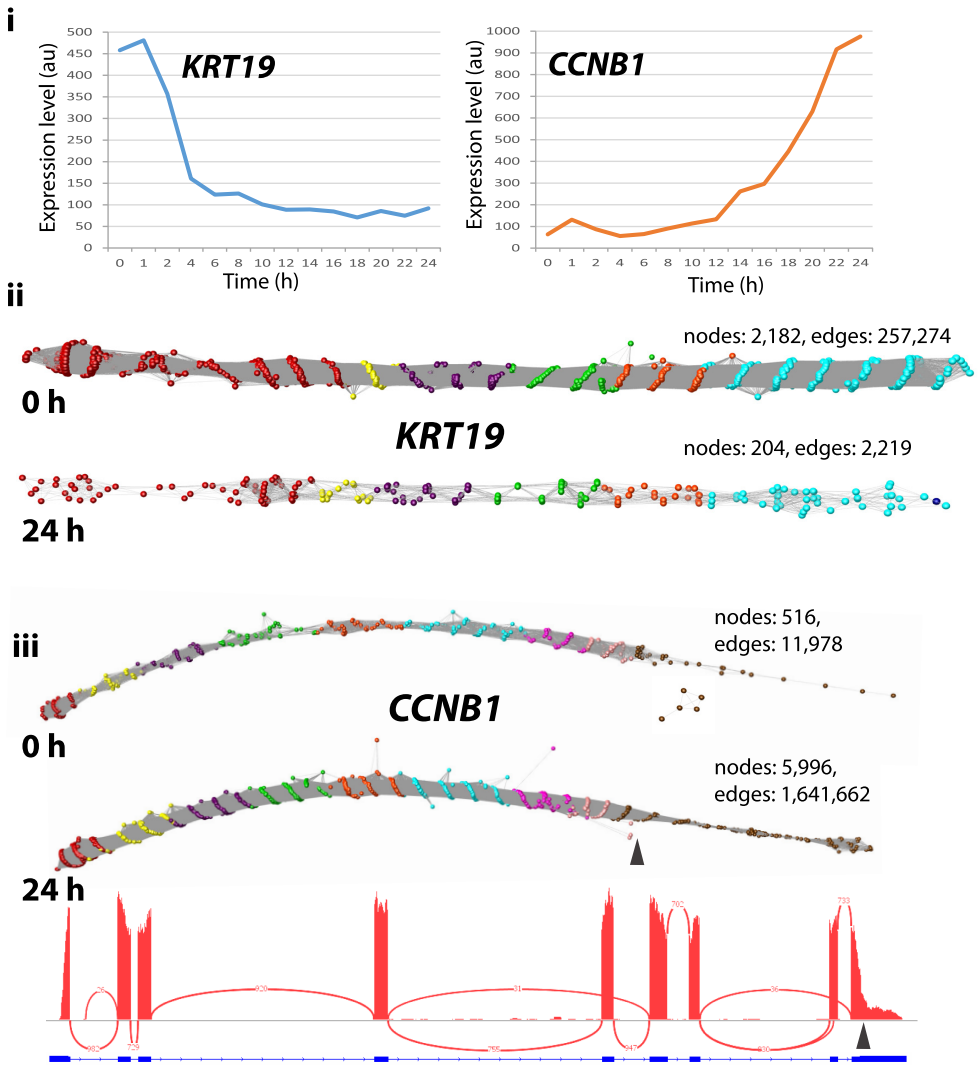

Figure 4. Typical graphs of RNA-Seq data derived from linear transcripts. Examination of a wide range of RNA-Seq assembly graphs derived from human fibroblast expressed genes demonstrated that the majority of graphs are linear structures. Shown here are two such graphs for **(ii)***KRT19* and **(iii)** *CCNB1* (node overlay colours representing exons were derived from ENST00000361566 and ENST00000256442, respectively). **(i)** Expression profile of the two genes as measured by microarray analysis of the time-course of transcriptional events following serum-refeeding (31). Expression of *KRT19* is rapidly down-regulated whilst *CCNB1* is up-regulated as the cells enter mitosis. This differential expression is evident from the graph visualization with the number of nodes decreasing or increasing by ∼10-fold in the 0 h derived versus the 24 h derived RNA-Seq data. It is interesting to note that in the *CCNB1* graphs there is an abrupt decrease in the density of nodes within exon 9 at both time points (marked by arrow). This corresponds to where the IGV view also shows a decrease in the density of reads and corresponds to a *CCNB1* transcript (ENST00000505500) that exhibits a truncated exon 9 at this position.

brain – 1006 reads) using public human RNA-Seq data (see Methods). In each case the RNA-assembly graph demonstrated varying degrees of complex topology (Figure 7A).

Graph analysis of *TPM1* expression in the heart revealed that there was essentially only one major transcript isoform expressed in this organ (ENST00000403994) containing exons 1a, 2b, 3, 4, 5, 6b, 7, 8 and 9a/b, the deep coverage of which can be inferred from the size of nodes. However, there was evidence of the presence of other minor transcript isoforms, visualized as small branches emanating from the core network. Mapping the reads from these minor structures to *TPM1* exons suggested the expression of up to four other minor transcript isoforms in the heart, based on the presence of isoform-specific exons. In contrast the *TPM1* assembly graph derived from liver suggests at least six isoforms expressed in this organ. The two major isoforms (ENST00000559556, ENST00000404484) possess alternative 5′ ends as can be observed from the bifurcation point at the 5′ end of the graph (Figure 7A). An alternative splicing event at exon 6 can also be inferred from the graph visualization and two 3′ termini are visible, both major isoforms ending in exon 9d and a minor form ending in exon 9a. There was also evidence of retained intronic sequence, shown by reads mapping to intron 6b. The most complex *TPM1* graph was derived from the brain. We observed what we believe to be eight isoforms expressed in brain tissue. There appear to be three major isoforms (ENST00000559556, ENST00000317516,

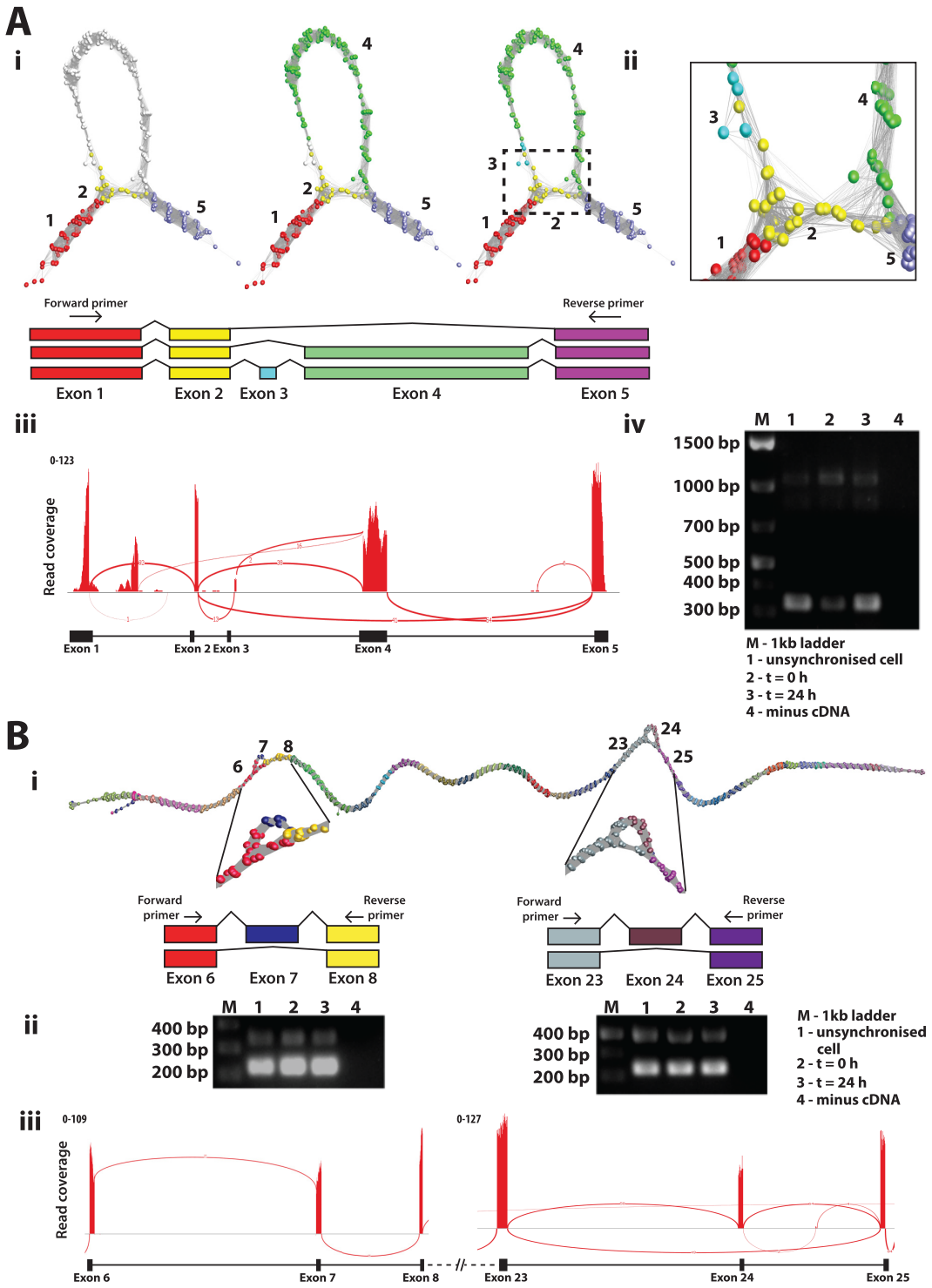

Figure 5. Splice variant visualization and confirmation. (**A**) Splice variants of *LRR1*. (**Ai**) Loop in the *LRR1* graph suggested that two or possibly three transcript isoforms are expressed in fibroblasts. These are shown overlaid on the graph together with a schematic representation of each isoform. (**Aii**) Close-up view of exon skipping event in *LRR1* shows the connection of exon 2 (yellow nodes) to exon 5 (light purple nodes) skipping exons 3 and 4 (blue/green nodes). (**Aiii**) Sashimi plot generated in IGV showing RNA-Seq reads mapping to the *LRR1* locus. Splice junctions are displayed as arcs connecting exons. The number of reads observed for each junction is indicated within segments, and y-axis ranges for the number of reads per exon base are shown. (**Aiv**) Result for RT-PCR of *LRR1* mRNA using the human fibroblast RNA from the proliferation time course. Three bands showing on the gel represent alternatively spliced products due to exon 3/4 skipping (310 bp), exon 3 skipping (1022 bp) or the full transcript (1130 bp). The position of the PCR primers is shown in Ai**.** (**B**) Splice variants of *PCM1*. (**Bi**) Two different splicing events for *PCM1* were evident from the graph visualization of this gene. (**Bii**) Result for RT-PCR of *PCM1* for two different locations of splice variant using the human fibroblast RNA from the proliferation time-course. Skipping of exon 7 resulted in a PCR fragment size of 223 bp, compared with 339 bp when included, and skipping of exon 25 resulted in a PCR fragment of 331 bp compared with 496 bp when included. (**Biii**) Representative Sashimi plot generated in IGV.

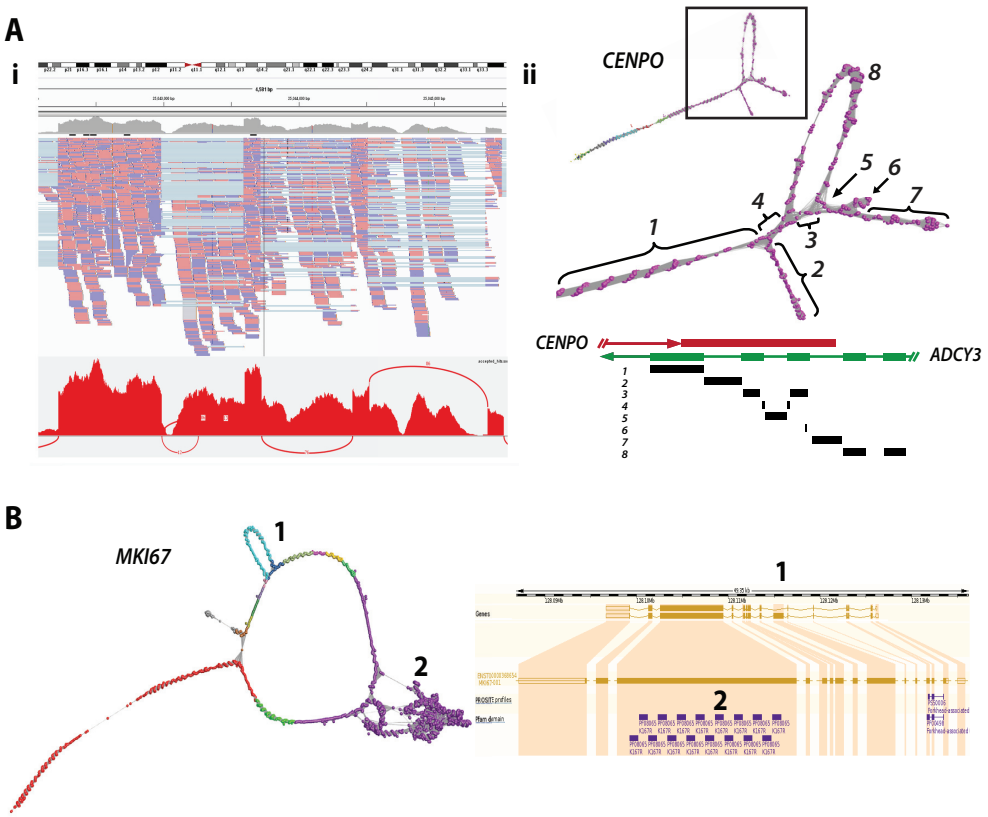

Figure 6. Complex gene graph structure. (**A**) Mis-assembly of reads from overlapping genes. (**Ai**) IGV visualization of *CENPO* exon 8 together with corresponding Sashimi plot. *ADCY3* overlaps with *CENPO* on the opposite strand of DNA. (**Aii**) In this case, reads derived from *ADCY3* mRNA are being wrongly mapped to *CENPO* resulting in the complex graph structure observed in exon 8. The schematic diagram of overlapping genes *CENPO* and *ACDY3* is shown above and regions of the graph mapped back to it. The loops in exon 8 of *CENPO* are formed by junction reads derived from *ADCY3* encoded on the opposite strand. (**B**) Repeat sequences cause perturbation in graph structure. Network-based visualization of *MKI67*. In this graph, there are two structures: (1) an alternatively spliced exon and (2) internal duplication. Skipping of exon 6 giving rise to ENST00000368653 can be observed as the loop structure, whilst the knotted structure is formed due the presence of 16 K167/Chmadrin repeat domains within exon 12.

ENST00000560975) and others showing evidence of retained introns. These retained intronic sequences can be inferred from the branching structures, e.g. introns 5, 6 and 8. The major splice isoforms are at the 3′ end at exon 9c and 9d. Another isoform can be inferred from the nodes branching at exon 8. Two isoforms in which intron 9a1 or exon 9b were retained were expressed in this sample. We compared the network-based analyses of *TPM1* as summarized in Figure 7C with the corresponding Sashimi plots (Figure 7D) and data from PacBio long-read sequencing of cDNA tissues from these tissues (34) shown as UCSC tracks (Figure 7E). In the case of the Sashimi plot we would argue that it is very difficult to work out the structure of the expressed transcripts from this visualization and whilst PacBio the long read result agrees with the graph analysis, it misses much of the complexity *TPM1* isoform expression, presumably due to low read depth. Finally, we looked at the expression of this gene in the fibroblast data. In this instance two isoforms of the transcript were observed, ENST00000358278 and ENST00000267996 which exhibited two mutually exclusive exon splicing events, the former containing exons 2b and 6a and the latter containing exons 2a and 6b (Figure 7F). The *TPM1* example explored here demonstrates something of the potential complexity of RNA-Seq data and where RNA assembly graphs may help with the interpretation expressed transcript isoforms.

DISCUSSION

RNA-Seq offers a platform to explore transcript diversity within and between cells and tissues. Sequencing platforms that produce short-read data (50–250 bp) currently dominate the field. Many tools and analysis pipelines already exist to process these data from the DNA sequencer, through mapping to a genome or *de novo* assembly, and summarize these data down to read counts per gene/transcript. These data are then ready for differential gene expression or cluster-based analyses. It is also routine practice to port data into tools such as IGV, where they can be visualized in the context of the reference genome (where available). Reads are shown stacked on to the region from which they have been determined to originate. In instances where a single RNA species is transcribed from a specific locus, existing visualizations are sufficient for most needs. However, where multiple transcripts are produced from a given locus, deconvolution of that assembly into the component transcripts can be challenging. Tools such as the Sashimi plot (22) use information derived from exon boundaries and paired end reads to display connections between exons, the thickness

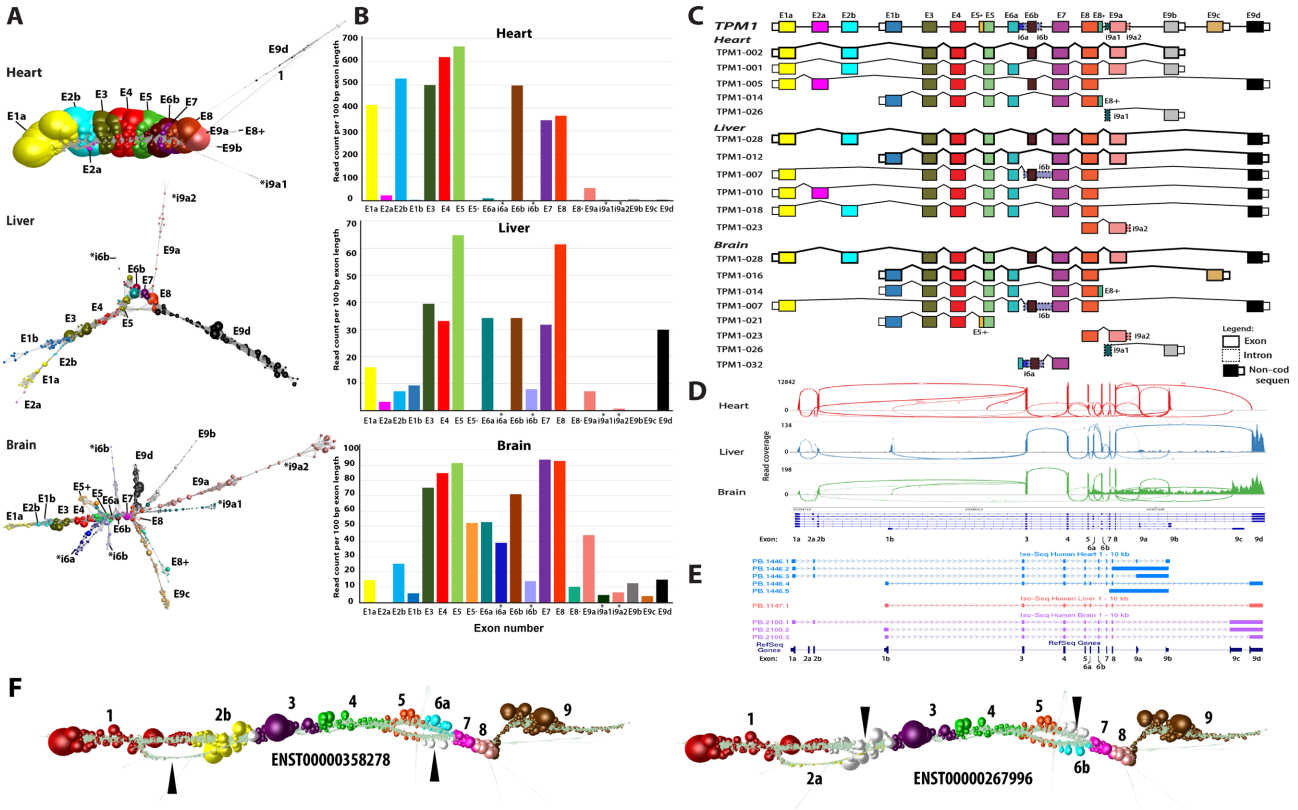

Figure 7. Graph visualization of *TPM1* transcript isoforms expressed in different tissues. (**A**) Graph-based visualization of *TPM1* mRNA from heart, liver and brain. In these graphs, node size reflects the relative read depth as shown in Figure 3. In the case of the heart, the *TPM1* expression graph shows one major isoform while an estimated four other minor isoforms are observed that give rise to network 'branches'. In the *TPM1* graph generated from liver, all isoforms would appear to be expressed at the similar levels. Alternative 5′ transcript initiation can be visualized as the first two branches of the graph. The mutually exclusive exons 6a and 6b can be seen from the graph structure, while the retention intron of 6b and 9a is observed as a few nodes branching out from the core graph structure. The RNA assembly graph generated from brain demonstrates that multiple isoforms are expressed in this tissue, a complex series of branched structures emanating out from the graph at different locations. The observed structure is supported by Ensembl gene models. (**B**) Histogram of the number of reads per 100 bp of exon per sample. (**C**) Schematic representation of *TPM1* transcript isoform diversity expressed in heart, liver and brain as interpreted from RNA assembly graphs. (**D**) Representative Sashimi plots generated in IGV showing RNA-Seq reads mapping to *TPM1* locus in the different tissues. Height of the bars represents read coverage. Splice junctions are displayed as arcs, with the number of reads observed for each junction being indicated by line thickness. (**E**) UCSC Genome Browser sequence visualization of *TPM1* gene from a different study of the whole human transcriptome of brain, heart and liver generated on the PacBio long-read platform. The number of isoforms detected by PacBio is different from the graph-based based visualization. PacBio detected five isoforms in heart (blue), one in liver (orange) and three in the brain (purple). (**F**) *TPM1* graph as expressed in fibroblasts. Two bifurcation events are clearly visible in the graph that correspond to two isoforms of the transcript, ENST00000358278 and ENST00000267996. These exhibit two mutually exclusive exon splicing events, the former containing exons 2b and 6a and the latter containing exons 2a and 6b, the white nodes (arrows) corresponding to reads that do not map to the corresponding transcript.

of the joining line indicating the number of reads that cross a given exon-exon boundary. When transcript diversity is relatively simple these views provide a sufficient representation of events, but when transcript variability is complex they can be difficult to interpret. Other tools for splice variant analysis focus on identifying statistically splice variants from data and whilst some also possess sophisticated visualizations, they generally rely of the read-stacking approach to display transcript isoforms and/or summary statistics of the number of reads that map to exons (17,49,50).

This work describes a complementary approach to the analysis and interpretation of RNA-Seq data, based on construction and visualization of RNA assembly graphs. In this method RNA-Seq reads mapping to a specific locus are directly compared with each other by calculating an all-vs-all similarity matrix. In the context of a graph visualization of these data, nodes represent individual reads or collections of identical reads, whilst edges represent similarity scores between them, above a given threshold. Information about read(s) can be used to annotate nodes, e.g. adjusting the size of a node to represent the number of reads it represents or colouring nodes based on the which exon they map to. In this manner the depth of sequencing and different transcript models can be overlaid easily on to the graph assemblies, reads derived from a given exon sharing the same colour. This provides a way to quickly visualize how well a given transcript model matches the structure of the assembly.

We demonstrate the ability to recognize alternative splicing events from these graphs, as well as areas with internal homology and issues with read mapping. A fundamental challenge in implementing this approach is the ability to layout and display graph assemblies of data, such that the underlying structure of the networks can be interpreted. Graphia Professional, formerly known as BioLay-

out *Express*^{3D}, was originally developed for the visualization and analysis gene expression data as large correlation networks (29,30) and pathway models (51). RNA assembly graphs have a fundamentally different graph topology to many graphs, with nodes (reads) generally only sharing edges with others in up or downstream positions of the genome and their unusual structure required us to incorporate the FMMM algorithm (52) into the tool. At its highest quality setting, the FMMM algorithm provides a linear layout of the string-like topology of RNA assembly graphs, allowing clear visualization of the underlying global structure of the graph, whilst at a local level nodes form a corkscrew-like structure (Supplementary Figure S1).

Studies of the synthetic data emphasize the importance of read depth and read similarity thresholds in visualizing graph structure. The experiments showed that when sufficient data is available and the similarity thresholds are set appropriately, even small variations in transcript isoforms can be readily observed (Figure 2). At the point where every base position along a transcript has the start of one read mapping to it, in principle further reads add nothing to an assembly graph's structure. Additional reads also add greatly to the computation time taken to calculate a read similarity matrix, and more nodes and edges have to be rendered. Collapsing redundant reads to a unique sequence/node therefore speeds up all aspects of visualization. Currently the analysis pipeline only simplifies down to unique reads, so reads that contain a polymorphism or sequencing errors are represented as separate nodes. Abundant identical reads are represented by large nodes, the size of the node decreasing with read depth. The graph assemblies for *GAPDH* and *TUBA1C* (Figure 3) are used to illustrate this approach. These graphs represent single transcripts sequenced at high depth, the many small nodes representing sequencing 'noise'. In order to minimize layout times and improve the fluidity of graph rendering, a similarity threshold should ideally be selected that allows the construction of a single component network with a maximum number of nodes and a minimum number of edges (Figure 3B). When a sole transcript is expressed, a linear string-like graph is generated. Up to a point, the relative depth of sequencing for a given transcript is immediately obvious from its graph visualization. In two cases shown of *KRT19* and *CCNB1* (Figure 4) the decrease and increase in their expression is reflected in the density of nodes in the resulting graphs, as well as the alternative splicing at the 3′ end of the latter.

Whenever an RNA assembly graph takes on higher order structure it is likely that sequences diverge, contain homologous domains, or may indicate an issue with sequence assembly. Where more than one transcript from the same gene is expressed in a sample, forks or loops in the graph are observed starting and finishing at exon boundaries, as demonstrated using the simulated data. *LRR1* as expressed in human fibroblasts is a relatively simple example of where two alternatively spliced transcripts are expressed by the same cell population. In one version of the *LRR1* transcript expressed by fibroblasts, exon 4 is spliced out and a large loop is observed in the graph (Figure 5A). The graph of *PCM1* possesses two loops corresponding to known splice variants at exons 7 and 24, both validated here. These splicing events are immediately obvious from the network visualizations, but perhaps less easy to appreciate from the corresponding sashimi plots (Figure 5B). In certain gene graphs we observed intra-exonic secondary structure. Within exon 8 of *CENPO* we observed complex network topology (Figure 6A). In this case our analyses showed that it was due to reads originating from *ADCY3* expression, whose terminal 5′ exons overlap on the opposite strand, causing loops within graph representing exon of *CENPO*. The inability to correctly map reads from over-lapping transcribed exons is one of the reasons the majority of RNA-Seq analyses are now generated from directional cDNA libraries. In the case of *MKI67* a series of 14 K167/Chmadrin domains present within exon 14 of the gene cause a knotted structure in that portion of the graph. An alternative splice variant missing exon 6 is also apparent in the graph visualization of this gene (Figure 6B).

We next wanted to test the potential of graph visualization to analyse transcript complexity within and between tissues. In the case shown here, we examined *TPM1* (tropomyosin 1) transcript diversity in RNA-Seq data derived from three human tissues, heart, liver and brain. *TPM1* is most highly expressed in the heart (and other muscles) where it functions as an actin-binding protein involved in the contractile system of muscles (53). Correspondingly, we observed a dominant and possibly sole functional transcript isoform expressed in heart, although a relatively small number of reads mapped to exon 2a and to terminal intronic sequences, suggesting the presence of a low number of other transcript isoforms. Whether these represent transcriptional 'noise' or transcription of other isoforms by cell types present in low abundance in the heart, is not clear. Expression levels of *TPM1* in the liver and brain are approximately 10 times lower than in the heart. In these tissues TPM1 is thought to play different roles in cytoskeletal organization (53). The corresponding RNA assembly graphs for *TPM1* in these tissues exhibited complex topologies. Through studying these graphs and mapping this information to the Ensembl transcript models for this gene, we estimated up to 6 transcript isoforms to be expressed in the liver, versus 10 in the brain. This is largely based on the presence in the data of reads mapping to transcript-specific exons.

It is reasonable to ask how much more information does graph visualization provide over and above of those of existing visualizations. We would view graphs, as generated here, as providing highly detailed visualizations of transcript complexity or in principle any assembly of reads. In some cases they provide a simple and unambiguous view of splice variation, as shown here for *LRR1* and *PCM1*, which are arguably simpler to interpret than alternative visualizations. In other cases transcript variation is inherently complex, as are the graphs, as illustrated by the example of *TPM1* (Figure 7A). In such circumstances there is no question that the graphs are challenging to interpret, but so are alternate views. From the sashimi plots it might appear that expression of transcript *TPM1* isoforms in the heart is as varied as in liver or brain, if not more so (Figure 7D). This is likely due to the much higher expression levels in the heart resulting reads associated with minor iso-

forms or transcriptional noise, giving rise to many apparent isoforms, as indicated exon-exon junction edges. In contrast, the dominance of one isoform expressed by the heart is clear from the RNA assembly graph. Where other approaches are clearly superior, are in contrasting expression of exons across samples and in providing a means to easily quickly explore multiple transcripts, something that at the current time is not possible using the graph-based approach described here. One limiting factor of this approach is the requirement to know in advance which portions of the data might be worth visualizing as a graph. Hence, we recommend that other tools are used to detect splice variants such as DEXSeq (54), Cuffdiff (55), JunctionSeq (19) or rMATs (49) prior to choosing genes for network analysis. Ultimately, long-read sequencing technologies, such as those produced Pacific Biosciences (PacBio) and Oxford Nanopore, will potentially circumvent the issue of identifying transcript isoforms in this way, as individual reads will encapsulate full length transcripts. However, in our analyses of *TPM1* transcript isoforms in the public PacBio data generated from the three tissues studied here (heart, brain and liver), there was only partial agreement with our graph-based analyses of corresponding short-read data. Although we detected all the isoforms identified by PacBio sequencing, many additional transcript isoforms were suggested by our analysis and not reported in the PacBio data, presumably due to a lack of coverage.

Whilst there are clearly limitations to the graph-based approaches, it can be envisaged that some of the current restrictions could be overcome. Assemblies could be further collapsed prior to visualization where, for example, nodes could represent small segments of a transcript or DNA sequence rather than individual reads and edges, the number of reads that overlap between these fragments. In this way graphs would be greatly simplified allowing many assemblies to be viewed at once and larger portions of sequencing data to be rendered. We are also currently working on tools where even larger graphs can be visualized and edge thresholds changed dynamically, enabling the graphs to filtered on-the-fly, supporting more rapid and detailed analyses of sequencing assemblies. In addition, we certainly see potential for this approach in analysing transcript assemblies where no reference genome is available and options for splice analysis are limited, or indeed for the analysis of any DNA assembly where higher order sequence complexities may be present. In summary, we believe the data pipeline, the tools and basic approach presented here provide an effective analytical paradigm that is a novel contribution to the analysis of the huge amounts of information-rich but complex data produced by modern DNA sequencing platforms.

DATA AVAILABILITY

Documentation and full source code for the NGS Graph Generator package can be downloaded from: https://github.com/systems-immunology-roslin-institute/ngs-graph-generator. *NGS graph generator*, a web interface for running the pipeline on pre-supplied datasets, can be accessed at http://seq-graph.roslin.ed.ac.uk. The network analysis tool Graphia Professional is free and open-source and can be downloaded here: https://kajeka.com/graphia-professional/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC) [BB/J019267/1]; Roslin Institute Strategic Grant from the Biotechnology and Biological Sciences Research Council (BBSRC) [BBS/E/D/10002071]; Majlis Amanah Rakyat (MARA), a Malaysian Government agency (to F.W.N.). Funding for open access charge: Biotechnology and Biological Sciences Research Council; Mater Research Institute - The University of Queensland receives funding from the Mater Foundation, Brisbane, Australia and the Translational Research Institute is supported by a grant from the Australian Government.

Conflict of interest statement. The analyses described here employ the network analysis software Graphia Professional distributed by Kajeka Ltd, Edinburgh. T.C.F. is a founder and director of Kajeka. The remaining authors have no conflict of interest.

REFERENCES

1. Morozova,O. and Marra,M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.
2. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
3. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
4. Yang,I.S. and Kim,S. (2015) Analysis of whole transcriptome sequencing data: Workflow and software. *Genomics Informatics*, **13**, 119–125.
5. Costa-Silva,J., Domingues,D. and Lopes,F.M. (2017) RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS ONE*, **12**, e0190152.
6. Han,Y., Gao,S., Muegge,K., Zhang,W. and Zhou,B. (2015) Advanced applications of RNA sequencing and challenges. *Bioinform. Biol. Insights*, **9**, 29–46.
7. Wang,J., Ye,Z., Huang,T.H., Shi,H. and Jin,V. (2015) A survey of computational methods in transcriptome-wide alternative splicing analysis. *Biomol. Concepts*, **6**, 59–66.
8. Pohl,M., Bortfeldt,R.H., Grutzmann,K. and Schuster,S. (2013) Alternative splicing of mutually exclusive exons–a review. *Bio. Syst.*, **114**, 31–38.
9. Bahrami-Samani,E., Vo,D.T., de Araujo,P.R., Vogel,C., Smith,A.D., Penalva,L.O. and Uren,P.J. (2015) Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput. *Wiley Interdiscipl. Rev. RNA*, **6**, 291–310.
10. Barann,M., Zimmer,R. and Birzele,F. (2017) Mananaggal - a novel viewer for alternative splicing events. *BMC Bioinformatics*, **18**, 120.
11. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.*, **14**, 178–192.
12. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
13. Milne,I., Bayer,M., Stephen,G., Cardle,L. and Marshall,D. (2016) Tablet: Visualizing next-generation sequence assemblies and mappings. *Methods Mol. Biol.*, **1374**, 253–268.

14. Carver,T., Harris,S.R., Otto,T.D., Berriman,M., Parkhill,J. and McQuillan,J.A. (2013) BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinform.*, **14**, 203–212.
15. Huang,W. and Marth,G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res*, **18**, 1538–1543.
16. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
17. Strobelt,H., Alsallakh,B., Botros,J., Peterson,B., Borowsky,M., Pfister,H. and Lex,A. (2016) Vials: visualizing alternative splicing of genes. *IEEE Trans. Vis. Comput. Graph.*, **22**, 399–408.
18. Liu,Q., Chen,C., Shen,E., Zhao,F., Sun,Z. and Wu,J. (2012) Detection, annotation and visualization of alternative splicing from RNA-Seq data with splicingviewer. *Genomics*, **99**, 178–182.
19. Hartley,S.W. and Mullikin,J.C. (2016) Detection and visualization of differential splicing in RNA-Seq data with Junction Seq. *Nucleic Acids Res.*, **44**, e127.
20. Ding,L.Z., Rath,E. and Bai,Y.S. (2017) Comparison of alternative splicing junction detection tools using RNA-Seq data. *Curr. Genomics*, **18**, 268–277.
21. Hooper,J.E. (2014) A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum. Genomic*, **8**, 3.
22. Katz,Y., Wang,E.T., Silterra,J., Schwartz,S., Wong,B., Thorvaldsdottir,H., Robinson,J.T., Mesirov,J.P., Airoldi,E.M. and Burge,C.B. (2015) Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, **31**, 2400–2402.
23. Lahat,A. and Grellscheid,S.N. (2016) In: Aransay,AM and Lavín,Trueba JL (eds). *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*. Springer International Publishing, Cham, pp. 105–119.
24. Novák,P., Neumann,P. and Macas,J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics*, **11**, 378.
25. Benoit-Pilven,C., Marchet,C., Chautard,E., Lima,L., Lambert,M.-P., Sacomoto,G., Rey,A., Cologne,A., Terrone,S., Dulaurier,L. *et al.* (2018) Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Scientific Rep.*, **8**, 4307.
26. Nielsen,C.B., Jackman,S.D., Birol,I. and Jones,S.J. (2009) ABySS-Explorer: visualizing genome sequence assemblies. *IEEE Trans. Vis. Comput. Graph.*, **15**, 881–888.
27. Wick,R.R., Schultz,M.B., Zobel,J. and Holt,K.E. (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350–3352.
28. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
29. Freeman,T.C., Goldovsky,L., Brosch,M., van Dongen,S., Mazière,P., Grocock,R.J., Freilich,S., Thornton,J. and Enright,A.J. (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.*, **3**, e206.
30. Theocharidis,A., van Dongen,S., Enright,A.J. and Freeman,T.C. (2009) Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat. Protoc.*, **4**, 1535–1550.
31. Giotti,B., Chen,S.H., Barnett,M.W., Regan,T., Ly,T., Wiemann,S., Hume,D.A. and Freeman,T.C. (2018) Assembly of a parts list of the human mitotic cell cycle machinery. *J. Mol. Cell Biol.*.doi:10.1093/jmcb/mjy063.
32. Davis,M.P., van Dongen,S., Abreu-Goodger,C., Bartonicek,N. and Enright,A.J. (2013) Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.
33. Fagerberg,L., Hallstrom,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics*, **13**, 397–406.
34. PACBIO (2014) Data Release: Whole Human Transcriptome from Brain, Heart, and Liver. *Pacific Biosci.*, https://www.pacb.com/blog/data-release-whole-human-transcriptome/.
35. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
36. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
37. Lawrence,M., Huber,W., Pages,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
38. Morgulis,A., Coulouris,G., Raytselis,Y., Madden,T.L., Agarwala,R. and Schaffer,A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.
39. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
40. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
41. Hachul,S. and Jünger,M. (2005) In: Pach,J (ed). *Graph Drawing: 12th International Symposium, GD 2004, New York, NY, USA, September 29-October 2, 2004, Revised Selected Papers*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 285–295.
42. Chimani,M., Gutwenger,C., Jünger,M., Klau,G.W., Klein,K. and Mutzel,P. (2013) The Open Graph Drawing Framework (OGDF). *Handb. Graph Draw. Visual.*, **2011**, 543–569.
43. Dobin,A. and Gingeras,T.R. (2015) Mapping RNA-seq reads with STAR. *Curr. Protoc. Bioinformatics*, **51**, 11.14.11–11.14.19.
44. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
45. Starostina,N.G., Simpliciano,J.M., McGuirk,M.A. and Kipreos,E.T. (2010) CRL2(LRR-1) targets a CDK inhibitor for cell cycle control in C. elegans and actin-based motility regulation in human cells. *Dev. Cell*, **19**, 753–764.
46. Wang,G., Chen,Q., Zhang,X.Y., Zhang,B.Y., Zhuo,X.L., Liu,J.J., Jiang,Q. and Zhang,C.M. (2013) PCM1 recruits Plk1 to the pericentriolar matrix to promote primary cilia disassembly before mitotic entry. *J. Cell Sci.*, **126**, 1355–1365.
47. Saito,A., Muro,Y., Sugiura,K., Sugiura,K., Ikeno,M., Ikeno,M., Yoda,K. and Tomita,Y. (2009) CENP-O, a protein localized at the centromere throughout the cell cycle, is a novel target antigen in systemic sclerosis. *J. Rheumatol.*, **36**, 781–786.
48. Takagi,M., Natsume,T., Kanemaki,M.T. and Imamoto,N. (2016) Perichromosomal protein Ki67 supports mitotic chromosome architecture. *Genes Cells*, **21**, 1113–1124.
49. Shen,S., Park,J.W., Lu,Z.-x., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5593–E5601.
50. Sun,W., Duan,T., Ye,P., Chen,K., Zhang,G., Lai,M. and Zhang,H. (2018) TSVdb: a web-tool for TCGA splicing variants analysis. *BMC Genomics*, **19**, 405.
51. O'Hara,L., Livigni,A., Theo,T., Boyer,B., Angus,T., Wright,D., Chen,S.H., Raza,S., Barnett,M.W., Digard,P. *et al.* (2016) Modelling the structure and dynamics of biological pathways. *PLoS Biol.*, **14**, e1002530.
52. Hachul,S. and Jünger,M. (2005) Drawing large graphs with a potential-field-based multilevel algorithm. In: Pach,J (ed). *Graph Drawing. GD 2004. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg , Vol. **3383**, pp. 285–295.
53. Perry,S.V. (2001) Vertebrate tropomyosin: Distribution, properties and function. *J. Muscle Res. Cell M*, **22**, 5–49.
54. Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.
55. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.